doi:10.1093/nar/gks386

SurvNet: a web server for identifying network-based biomarkers that most correlate with patient survival data

Jun Li[1,2], Paul Roebuck[3], Stefan Grünewald[1] and Han Liang[2,*]

[1]Chinese Academy of Sciences Key Laboratory of Computational Biology, Chinese Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, P.R. China, [2]Department of Bioinformatics and Computational Biology and [3]Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Received March 2, 2012; Revised April 11, 2012; Accepted April 14, 2012

ABSTRACT

An important task in biomedical research is identifying biomarkers that correlate with patient clinical data, and these biomarkers then provide a critical foundation for the diagnosis and treatment of disease. Conventionally, such an analysis is based on individual genes, but the results are often noisy and difficult to interpret. Using a biological network as the searching platform, network-based biomarkers are expected to be more robust and provide deep insights into the molecular mechanisms of disease. We have developed a novel bioinformatics web server for identifying network-based biomarkers that most correlate with patient survival data, *SurvNet*. The web server takes three input files: one biological network file, representing a gene regulatory or protein interaction network; one molecular profiling file, containing any type of gene- or protein-centred high-throughput biological data (e.g. microarray expression data or DNA methylation data); and one patient survival data file (e.g. patients' progression-free survival data). Given user-defined parameters, SurvNet will automatically search for subnetworks that most correlate with the observed patient survival data. As the output, SurvNet will generate a list of network biomarkers and display them through a user-friendly interface. SurvNet can be accessed at http://bioinformatics.mdanderson.org/main/SurvNet.

INTRODUCTION

With the advance of genome characterization technology, high-throughput genomic and proteomic data of patients have accumulated rapidly, allowing the systematic identification of biomarkers (1–4). Biomarkers that correlate with patient survival data are of particular interest because they provide a critical foundation for the diagnosis and treatment of disease (5,6). Conventionally, such an analysis is based on individual genes. However, the results thereby obtained are often noisy and difficult to interpret the underlying mechanisms of disease. Biological networks (e.g. gene regulatory networks or protein interaction networks) represent a reasonable way to summarize the functional behaviours of components within a biological system (7–9). Therefore, using a biological network as the searching platform, network-based biomarkers (i.e. a group of functionally related genes or proteins) are expected to be more robust and provide valuable insights into the molecular mechanisms of disease. Previous studies (10,11) on this topic have focused on other clinical data (such as metastasis status), and the utility of patient survival data has not been explored.

In this study, we introduce SurvNet, a novel bioinformatics web server for identifying network-based biomarkers that most correlate with patient survival data. The web server takes three input files: one biological network file, one molecular profiling file and one patient survival data file. In order to identify network-based biomarkers, SurvNet uses established algorithms (10–12) for searching and evaluating the biomarkers. As the output, SurvNet generates a user-friendly display of network-based biomarkers. We expect SurvNet to be a valuable bioinformatics tool for the biomedical community.

MATERIALS AND METHODS

The computational approach used by SurvNet to identify network-based biomarkers consists of three component processes: (i) a scoring function (combining the

*To whom correspondence should be addressed. Tel: +1 713 745 9815; Fax: +1 713 563 4242; Email: hliang1@mdanderson.org subnetwork property, molecular profile and patient survival data), (ii) a searching algorithm (for finding the candidate biomarkers) and (iii) an evaluation (validating the statistical significance of the biomarkers).

Scoring function

SurvNet first evaluates each gene (node) i by calculating the P-values p_i from a univariable Cox proportional hazards regression model (13,14), which quantifies how significantly the molecular profiling data of the gene correlate with the patient survival data. Then, each gene i is assigned a z-score s_i transformed from p_i,

$$s_i = \Phi^{-1}(1 - p_i),$$

as the score for each node in the network, where Φ^{-1} is the inverse standard normal cumulative distribution function (12). For random data, p_i follows a uniform distribution from 0 to 1, and by the transformation, s_i follows a standard normal distribution, with smaller p_i corresponding to larger z-scores.

The scoring function F of a subnetwork G with n genes is calculated by an aggregate z-score (12),

$$F_G = \frac{1}{\sqrt{n}} \sum_{i \in G} s_i,$$

where F_G follows a standard normal distribution if the s_i are independently drawn from a standard normal distribution. According to the formula, F_G is independent with a subnetwork of size n. Therefore, subnetworks with different sizes are comparable under this score function.

Searching algorithm

Because finding the connected subnetworks with the maximal score is NP-hard (12), SurvNet uses a greedy searching algorithm, as previously described (10–12). The searching starts from a seeded gene i and expands iteratively. The algorithm will terminate and output the candidate subnetworks if no candidate gene j around the current subnetwork G satisfies the following two conditions: (i) the number of edges in the shortest path between j and seeded gene i is smaller than or equal to δ and (ii) the score of subnetwork G with gene j is higher than $(1 + \rho)$ * F_G, where δ and ρ are two pre-determined parameters. Specifically, δ is used to reduce the searching space and ρ is a fixed increasing rate, ensuring that a new gene added to the subnetwork must increase the network score F_G by a rate larger than or equal to ρ.

Evaluation

SurvNet evaluates the statistical significance of the subnetworks identified in the searching step, as previously described (12). It first uses random sampling to see if the score of a subnetwork is significantly higher than that of a random gene set in the network. To do so, SurvNet randomly samples gene sets with n genes 10 000 times. Then, the same scoring function is used to calculate the scores for the random gene sets. The population mean μ_n and standard deviation σ_n are estimated from the sampled gene sets. Finally, F_G is calibrated against this background distribution as follows:

$$\tilde{F}_{\mathrm{G}} = \frac{F_{\mathrm{G}} - \mu_{\mathrm{n}}}{\sigma_{\mathrm{n}}}.$$

This calibrated score is the final network score for a subnetwork in the output. Moreover, since the multivariable Cox proportional hazards regression model is widely used to quantify the correlation between a group of genes and patient survival data, SurvNet also calculates the mutivariable Cox P-values for each subnetwork to validate their clinical utility. One potential advantage of SurvNet is to identify key disease genes that could have been missed through single-gene based analyses. For example, TP53 is a master cancer gene in ovarian carcinoma. Based on the protein expression and patient survival data from a recent study (1), TP53 protein, as a single node, shows no significant correlation with the patient survival, but a TP53-centered network is among the top biomarkers SurvNet identifies.

WEB SERVER

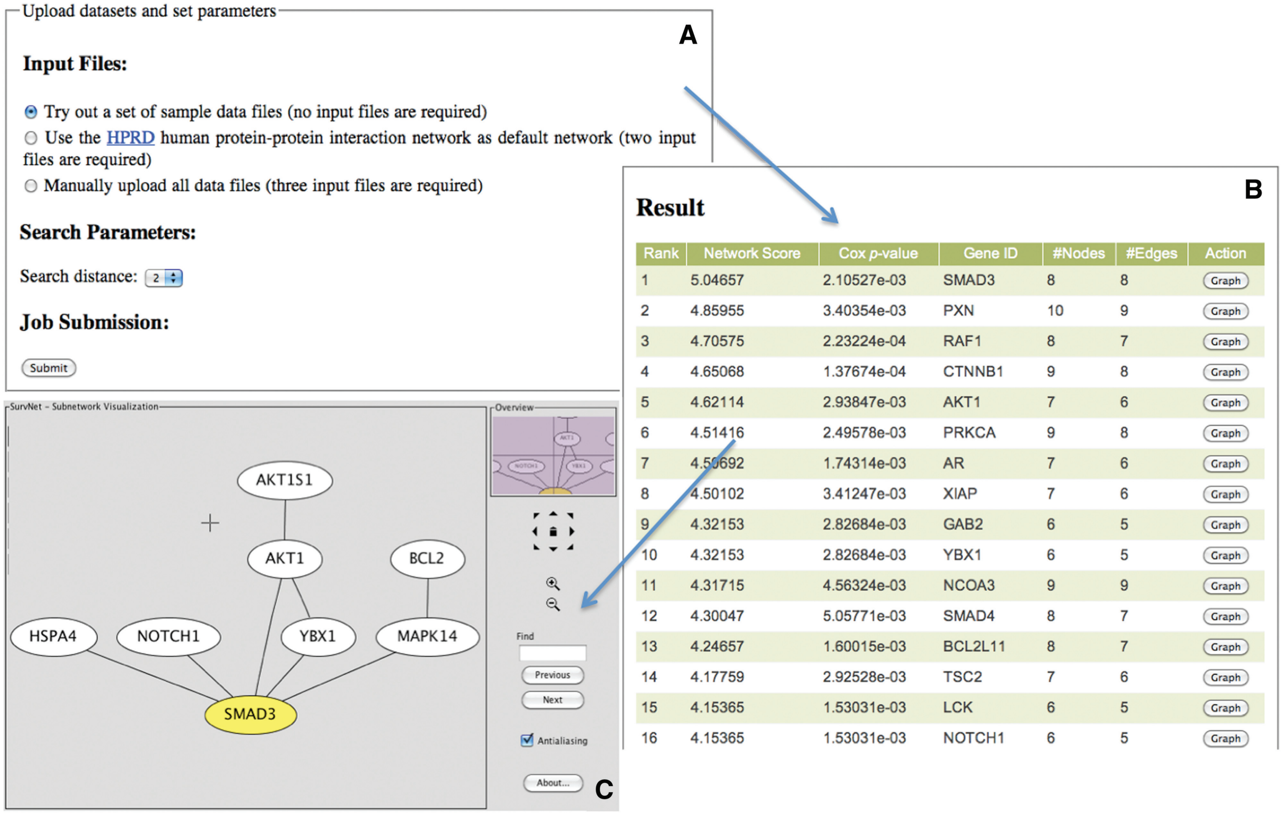
Input

The web server accepts three input files. The first one is a biological network file, representing a gene regulatory or protein interaction network [a human protein–protein interaction network (15) is provided as the default]. This file contains all the edges of a biological network, in which each line represents an edge. The second file is one molecular profiling file, containing any type of gene- or protein-centred high-throughput biological data (e.g. microarray-based gene expression data, reverse-phase protein array (16) protein expression data, DNA methylation data or gene mutation data). This file is a tab-separated numeric matrix, where the column names are the sample IDs and row names are gene IDs. The third file is one patient survival data file (e.g. patients' overall survival time or progression-free survival time). This file has three columns, named 'id', 'censor' and 'time', respectively.

After uploading the required input files, users can set the search distance (Figure 1A). This parameter defines the searching area in the network: start with each valid gene (or protein) node as the seed, SurvNet will automatically search for the optimal subnetwork(s) within this defined distance. SurvNet uses the same network searching algorithm that was previously described (10,11). A larger parameter will require a longer computation time. The default search distance is 2.

Output

In the final output, the subnetworks that SurvNet identifies will first be displayed in a table format (Figure 1B). These results can be directly downloaded. The network files are in a '.dot' format that can be visualized by GraphViz (http://www.graphviz.org). As shown in Figure 1C, the identified subnetworks are ranked within the table according to the network score,



Upload datasets and set parameters

Input Files:

● Try out a set of sample data files (no input files are required)
○ Use the HPRD human protein-protein interaction network as default network (two input files are required)
○ Manually upload all data files (three input files are required)

Search Parameters:

Search distance: 2

Job Submission:

Result

Rank	Network Score	Cox *p*-value	Gene ID	#Nodes	#Edges	Action
1	5.04657	2.10527e-03	SMAD3	8	8	Graph
2	4.85955	3.40354e-03	PXN	10	9	Graph
3	4.70575	2.23224e-04	RAF1	8	7	Graph
4	4.65068	1.37674e-04	CTNNB1	9	8	Graph
5	4.62114	2.93847e-03	AKT1	7	6	Graph
6	4.51416	2.49578e-03	PRKCA	9	8	Graph
7	4.50692	1.74314e-03	AR	7	6	Graph
8	4.50102	3.41247e-03	XIAP	7	6	Graph
9	4.32153	2.82684e-03	GAB2	6	5	Graph
10	4.32153	2.82684e-03	YBX1	6	5	Graph
11	4.31715	4.56324e-03	NCOA3	9	9	Graph
12	4.30047	5.05771e-03	SMAD4	8	7	Graph
13	4.24657	1.60015e-03	BCL2L11	8	7	Graph
14	4.17759	2.92528e-03	TSC2	7	6	Graph
15	4.15365	1.53031e-03	LCK	6	5	Graph
16	4.15365	1.53031e-03	NOTCH1	6	5	Graph

Figure 1. Snapshots of the SurvNet web server. (**A**) Input page, through which input files and a search parameter can be specified. (**B**) Output page, on which the top subnetwork biomarkers identified are displayed. (**C**) Visualization page, on which subnetworks can be visualized in a user-friendly way.

from high to low score. Each score is associated with the output items that follow. The network score, which quantifies how significantly the nodes in a subnetwork correlate with the observed patient survival data, is calculated based on the univariable Cox proportional hazards model and the network properties. A higher network score indicates a more significant correlation between the network and the patient survival time. The 'Cox *P*-value' is the *P*-value derived from the multivariable Cox proportional hazards regression model. The 'gene_ID' indicates the seeded node for each subnetwork; the number of nodes indicates how many genes (or proteins) are in the subnetwork; and the number of edges indicates how many interactions are in the subnetwork. Users can further narrow down the results by two output parameters: network *P*-value cut-off and minimal number nodes. The network *P*-value cut-off determines how significant the returned subnetworks are compared to the random background; and the default significance level is 0.05. Minimal number nodes determine the minimal number of nodes in a subnetwork; the default value is 2.

After clicking the 'Graph' button in the final output page, users can visualize an identified subnetwork in a user-friendly Java applet that allows them to pan/zoom, search and retrieve useful information (from GeneCard) for a node of interest. A detailed description about the Java applet is available under the visualization page.

CONCLUSION

We have developed SurvNet, a web server that can efficiently identify network-based biomarkers that most correlate with patient survival data. To the best of our knowledge, SurvNet is the only available bioinformatics tool for this function. SurvNet uses the network-based biomarker searching algorithms that were established in previous studies, and provides a user-friendly interface for exploring the identified biomarkers. We expect SurvNet to be a valuable resource for generating meaningful hypotheses for disease diagnosis and treatment.

ACKNOWLEDGEMENTS

We thank Bradley Broom, Jun Zhang, Allen Chang, Chris Wakefield, Glen Brosch and Sylvain Laroche for their help in building and maintaining the web server.

FUNDING

The University of Texas MD Anderson Cancer Center, the National Institutes of Health [U24 CA143883 and P30 CA016672]; G.S. Hogan Gastrointestinal Research fund and the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine (to H.L.). Funding for open access charge: The University of Texas MD Anderson Cancer Center.

Conflict of interest statement. None declared.

REFERENCES

1. Bell,D., Berchuck,A., Birrer,M., Chien,J., Cramer,D.W., Dao,F., Dhir,R., DiSaia,P., Gabra,H., Glenn,P. *et al.* (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.
2. Kalinina,J., Peng,J., Ritchie,J.C. and Van Meir,E.G. (2011) Proteomics of gliomas: initial biomarker discovery and evolution of technology. *Neuro Oncol.*, **13**, 926–942.
3. Chin,L., Meyerson,M., Aldape,K., Bigner,D., Mikkelsen,T., VandenBerg,S., Kahn,A., Penny,R., Ferguson,M.L., Gerhard,D.S. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
4. Schreiber,S.L., Shamji,A.F., Clemons,P.A., Hon,C., Koehler,A.N., Munoz,B., Palmer,M., Stern,A.M., Wagner,B.K., Powers,S. *et al.* (2010) Towards patient-based cancer therapeutics. *Nat. Biotechnol.*, **28**, 904–906.
5. Bast,R.C. Jr (2012) Molecular approaches to personalizing management of ovarian cancer. *Ann. Oncol.*, **22(Suppl. 8)**, viii5–viii15.
6. Mok,T.S.K. (2011) Personalized medicine in lung cancer: what we need to know. *Nat. Rev. Clin. Oncol.*, **8**, 661–668.
7. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
8. Kim,J. and Tan,K. (2010) Discover protein complexes in protein-protein interaction networks using parametric local modularity. *BMC Bioinformatics*, **11**, 521.
9. Voevodski,K., Teng,S.H. and Xia,Y. (2009) Finding local communities in protein networks. *BMC Bioinformatics*, **10**, 297.
10. Chuang,H.Y., Lee,E., Liu,Y.T., Lee,D. and Ideker,T. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.
11. Jia,P., Zheng,S., Long,J., Zheng,W. and Zhao,Z. (2011) dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, **27**, 95–102.
12. Ideker,T., Ozier,O., Schwikowski,B. and Siegel,A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18(Suppl. 1)**, S233–S240.
13. Cox,D.R. and Oakes,D. (1984) *Analysis of Survival Data*. Chapman and Hall, London, New York.
14. Hosmer,D.W. and Lemeshow,S. (1999) *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley, New York.
15. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
16. Tibes,R., Qiu,Y.H., Hennessy,B., Andreeff,M., Miiis,G.B. and Kornblau,S.M. (2006) Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol. Cancer Ther.*, **5**, 2512–2521.